eLIFE elifesciences.org

RESEARCH ARTICLE

Figures and figure supplements

Discovery of several thousand highly diverse circular DNA viruses

Michael J Tisza *et al*

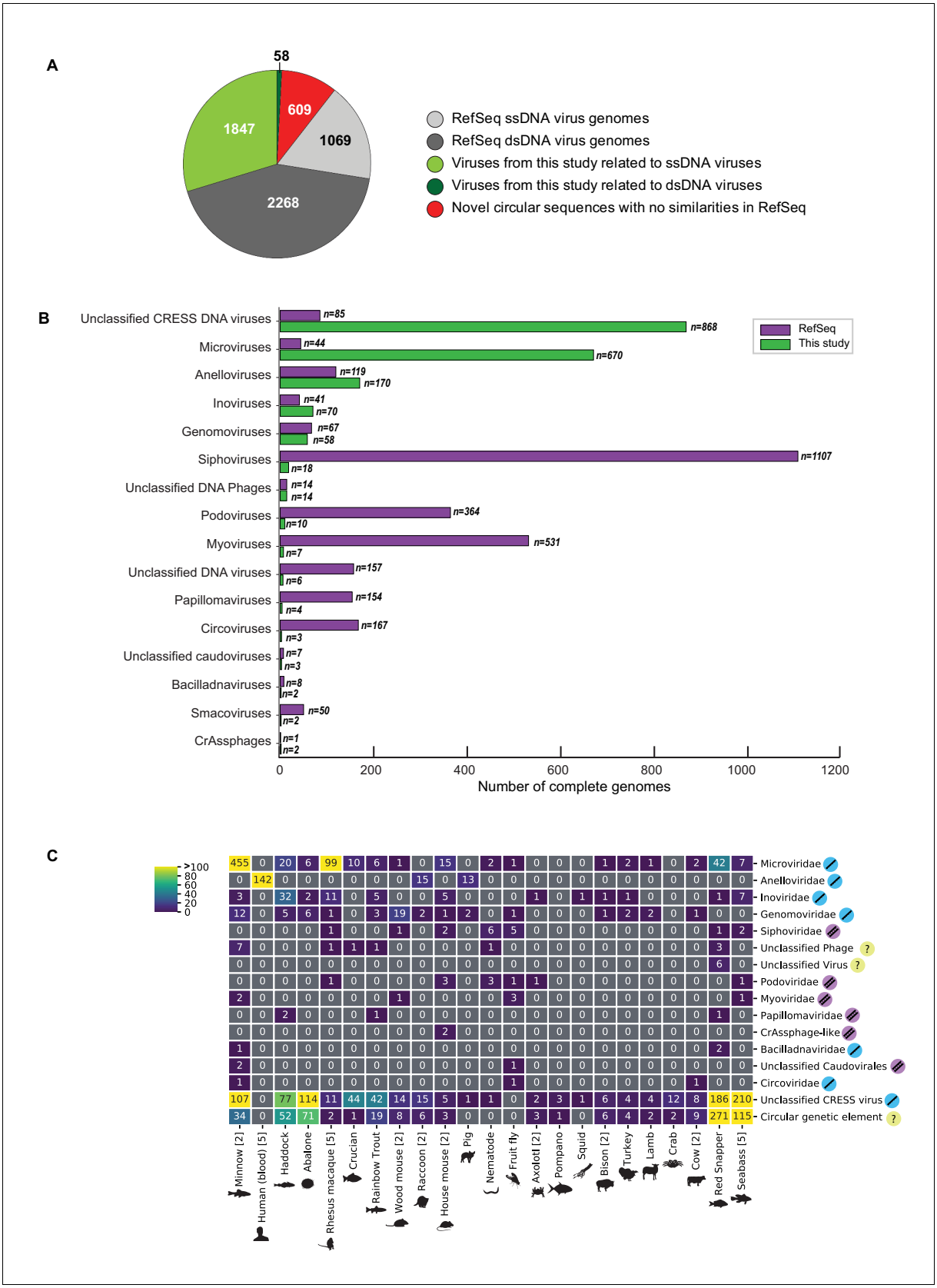

Figure 1. Novel viruses associated with animal samples. Gross characterization of viruses discovered in this project compared to NCBI RefSeq virus database entries. (**A**) Pie chart representing the number of viral genomes in broad categories. (**B**) Bar graph showing the number of new representatives

Figure 1 continued on next page

Figure 1 continued

of known viral families or unclassified groups. (**C**) Heatmap reporting number of genomes found associated with each animal species. Number of samples per species in brackets. Note that genomes in this study were assigned taxonomy based on at least one region with a BLASTX hit with an E value $<1 \times 10^{-5}$, suggesting commonality with a known viral family. Some genomes may ultimately be characterized as being basal to the assigned family.

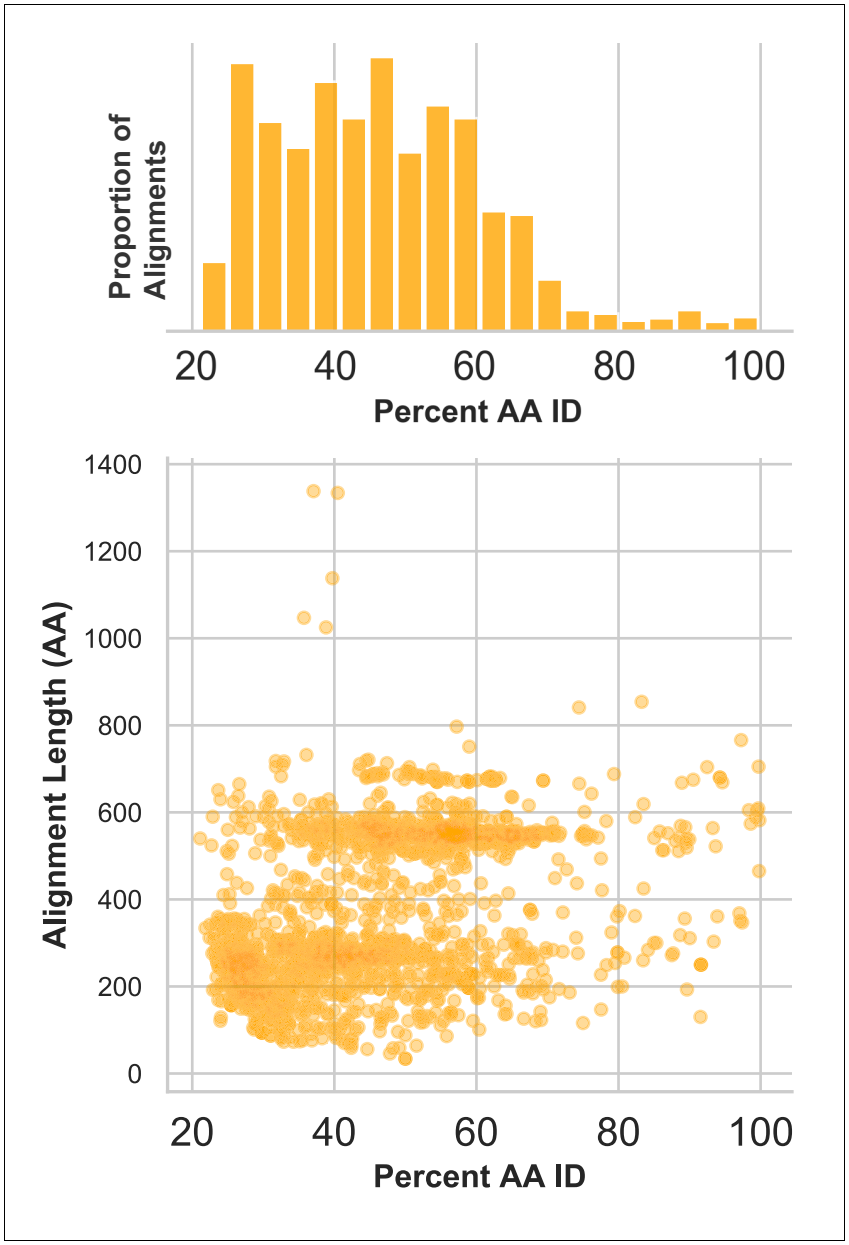

Figure 1—figure supplement 1. Divergence of proteins encoded by circular contigs. BLASTX summary of each circular DNA molecule recovered from virus-enriched samples. Sequences were queried against a database of viral and plasmid sequences. Only hits with E values < 10^{-5} were plotted. Here, BLASTX only reports the most significant stretch of amino acid sequence from each circular contig, and, therefore, other regions of each contig can be assumed to be equally or less conserved.

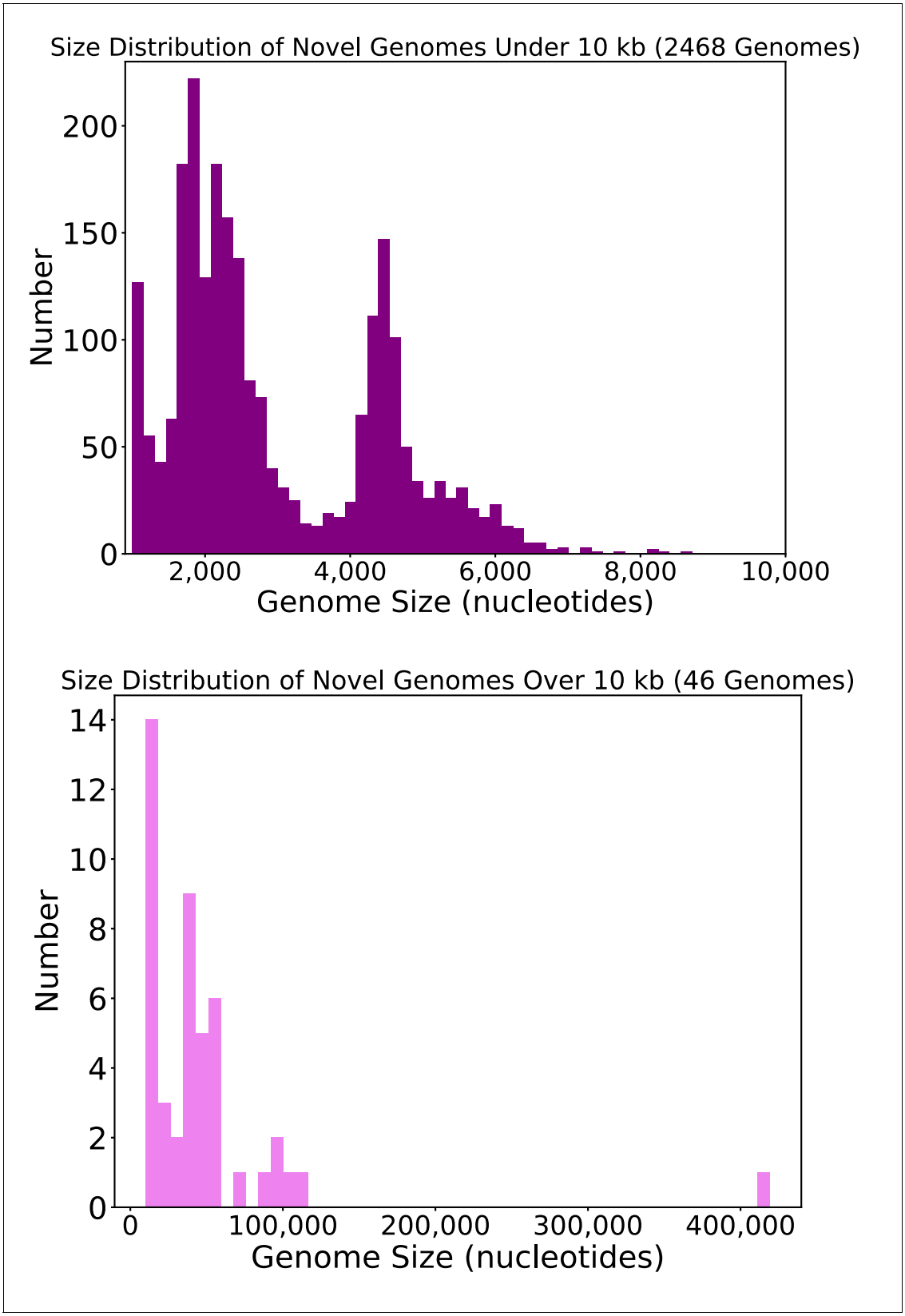

Figure 1—figure supplement 2. Size distribution of circular DNA sequences from this study. Length, in nucleotides, of circular DNA sequences representing putative viral genomes from this study.

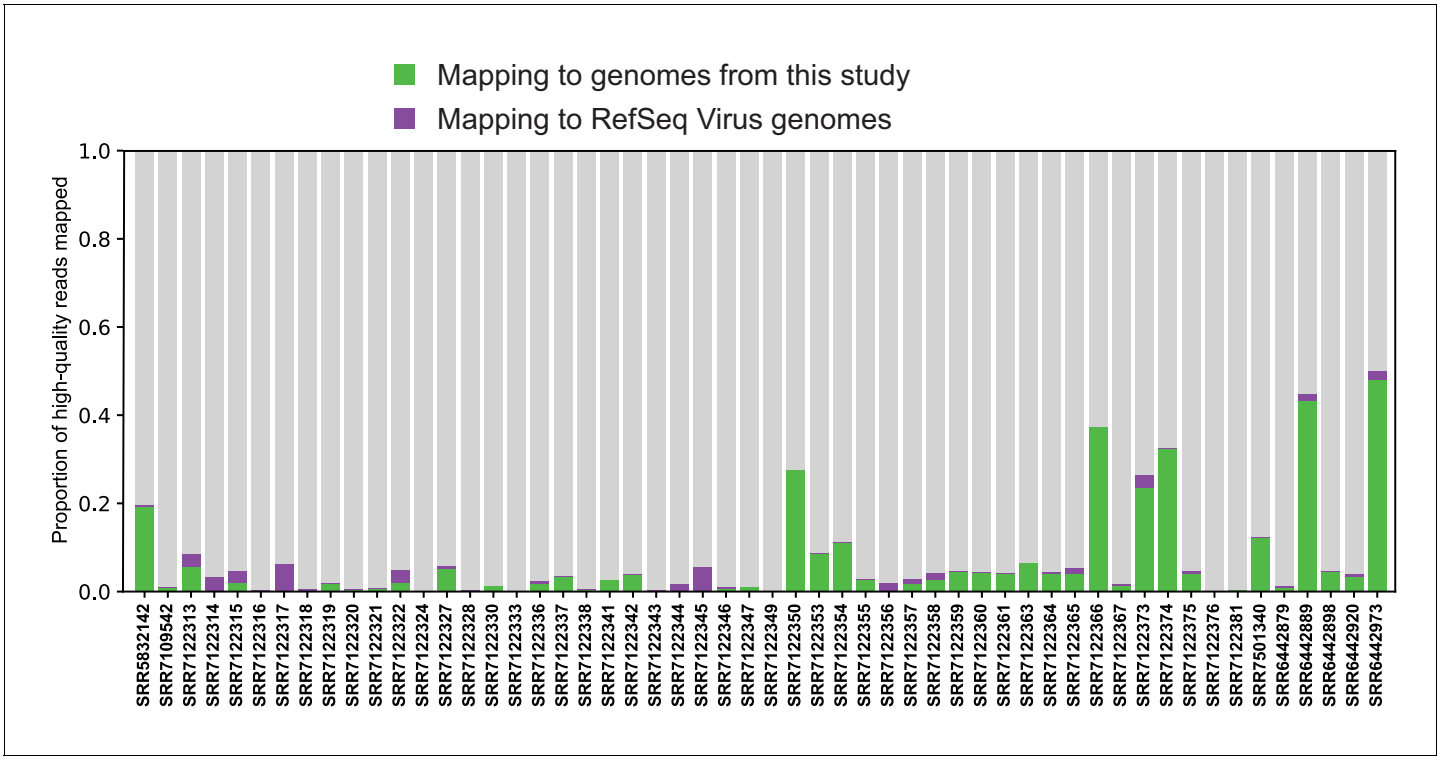

Figure 1—figure supplement 3. Mapping reads to complete viral genome references. Quality-trimmed reads were aligned with Bowtie2 to reference genomes from RefSeq and this study. Genomes were masked for low-complexity regions.

Figure 2. Sequence similarity network analysis of CRESS virus capsid proteins. EFI-EST was used to conduct pairwise alignments of amino acid sequences from this study and GenBank with predicted structural similarity to CRESS virus capsid proteins. The E value cutoff for the analysis was 10^{-5}.

Figure 2 continued on next page

Figure 2 continued

(**A**) Cluster consisting of proteins with predicted structural similarity to geminivirus-like capsids and/or STNV-like capsids. The phylogenetic tree was made from all sequences in this cluster. (**B**) A cluster consisting of sequences with predicted structural similarity to Circovirus capsid proteins. The phylogenetic tree was made from all sequences in this cluster. (**C**) Assorted clusters and singletons from unclassified CRESS virus proteins that were modeled to be capsids. (**D**) Nanovirus capsids. (**E**) Gyrovirus capsids.

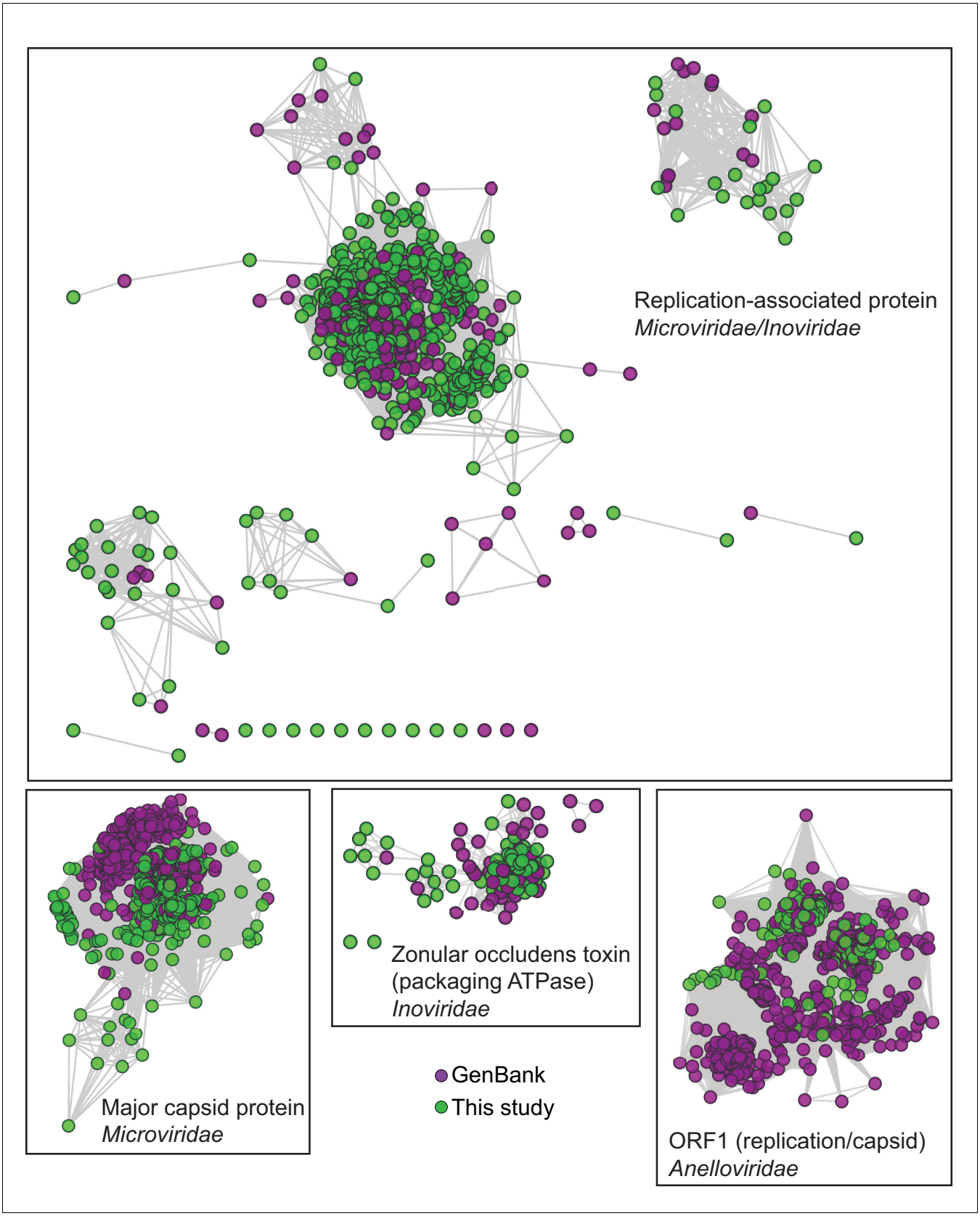

Figure 2—figure supplement 1. Network Analysis of additional viral hallmark genes. Depiction of additional viral hallmark genes from this study and GenBank as sequence similarity networks. E value cutoff = 10^{-5}. See ***Figure 2*** and Materials and methods.

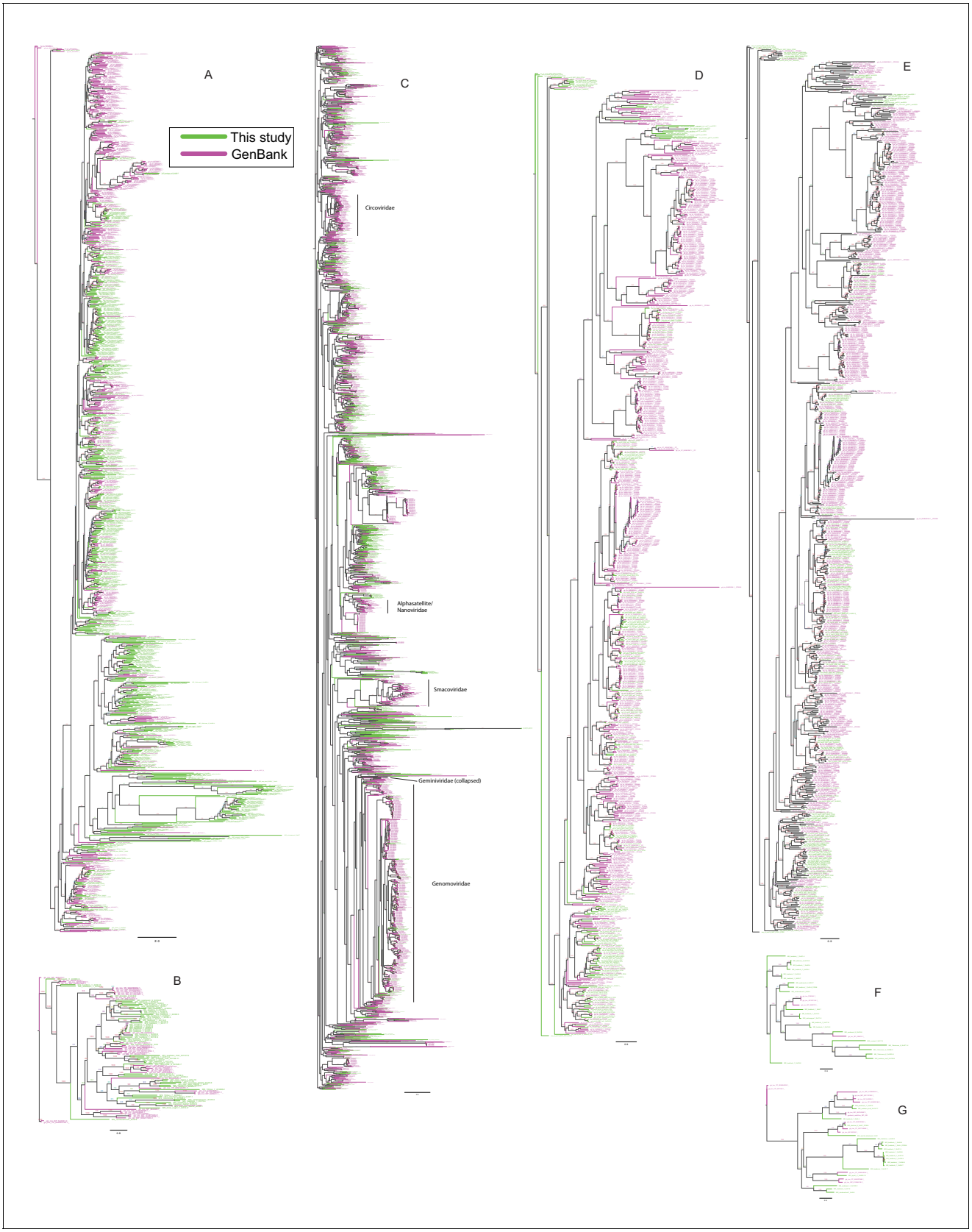

Figure 2—figure supplement 2. Phylogenetic trees of viral hallmark genes. Sequences were aligned with PROMALS3D using structure guidance when possible. Trees were drawn using IQ-Tree with automatic determination of substitution model. See Materials and methods. Branches are labeled with

Figure 2—figure supplement 2 continued on next page

Figure 2—figure supplement 2 continued

bootstrap percent support after 1000 ultrafast bootstrapping events. (**A**) *Microviridae* major capsid protein. (**B**) *Inoviridae* zonular occludens toxin. (**C**) CRESS virus Rep. (**D**) *Anelloviridae* ORF1 (**E**) *Microviridae*/*Inoviridae* Replication-associated protein I. (**F**) *Microviridae*/*Inoviridae* Replication-associated protein II. (**G**) *Microviridae*/*Inoviridae* Replication-associated protein III.

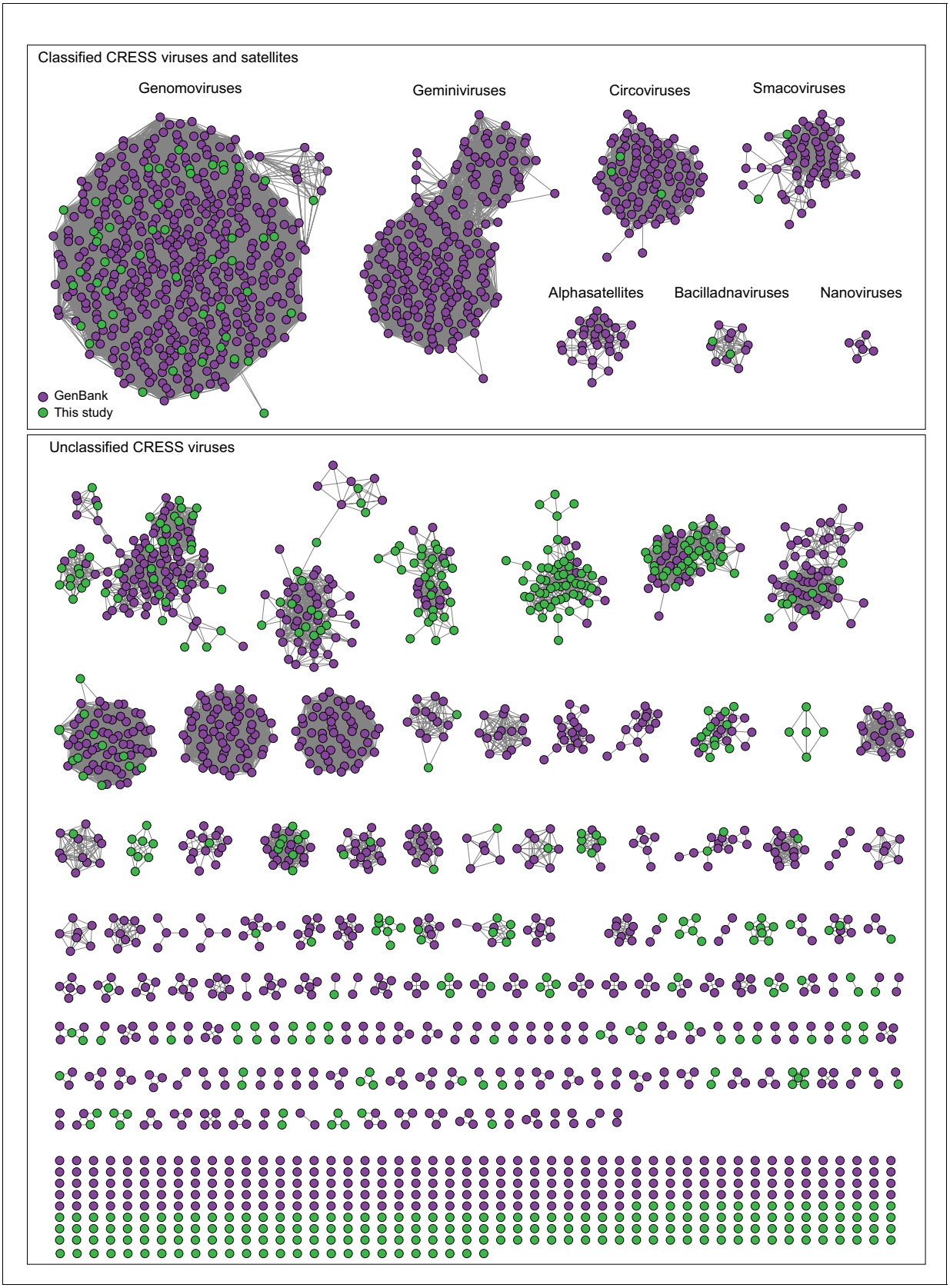

Figure 3. Network analysis of CRESS virus Rep proteins. EFI-EST was used to conduct pairwise alignments of amino acid sequences from this study and. GenBank that were structurally modeled to be a rolling-circle replicase (Rep). The analysis used an E value cutoff of 10^{-60} to divide the data into family-level clusters.

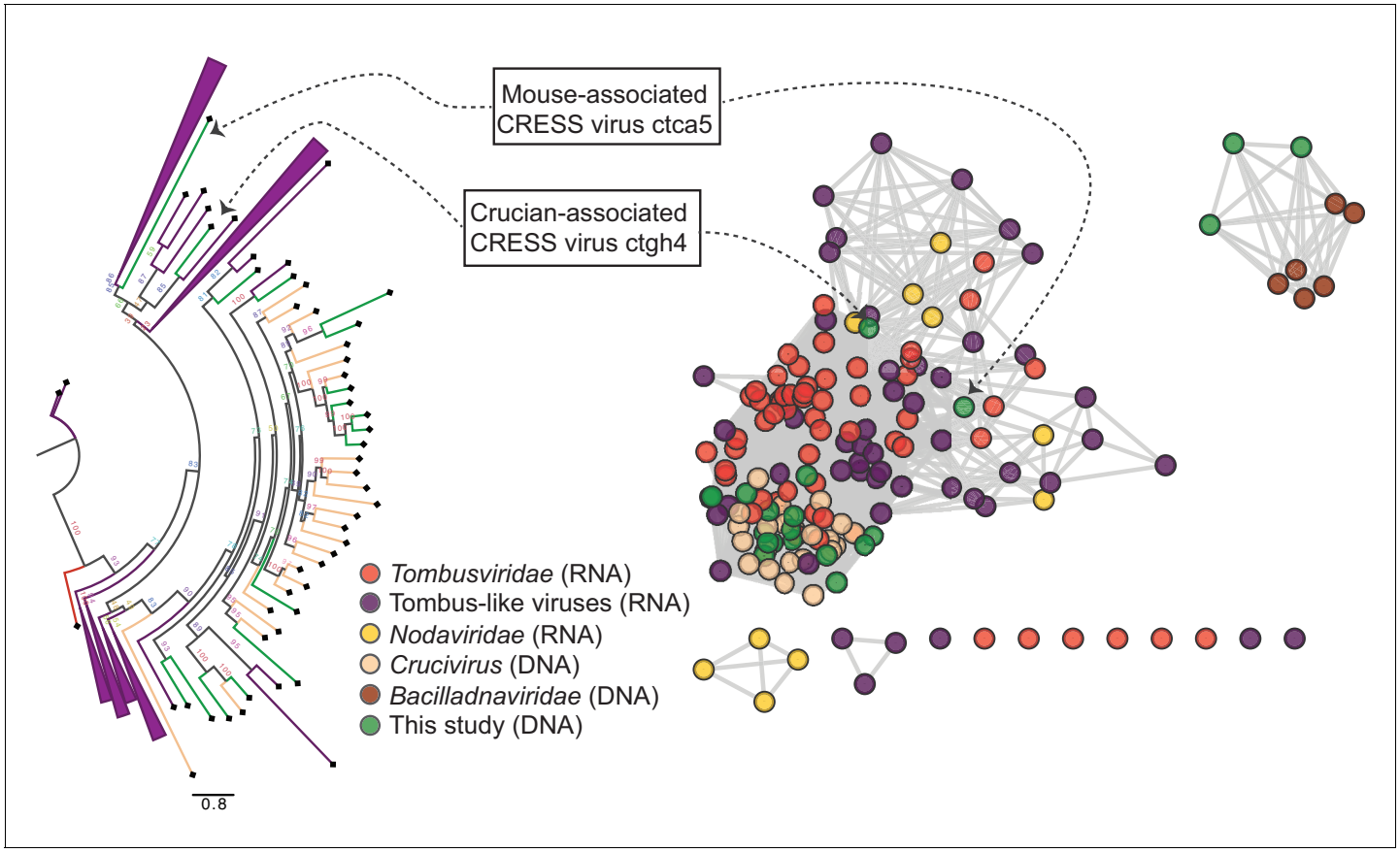

Figure 4. RNA virus capsid-like proteins. Sequence similarity network generated with EFI-EST (E value cutoff of 10^{-5}) showing capsid protein sequences of select ssRNA viruses (*Nodaviridae*, *Tombusviridae*, tombus-like viruses) and ssDNA viruses (*Bacilladnaviridae* and crucivirus) together with protein sequences from DNA virus genomes observed in the present study with predicted structural similarity to an RNA virus capsid protein domain (PDB: 2IZW). Predicted capsid proteins for CRESS virus ctca5 and CRESS virus ctgh4 have no detectable similarity to any known DNA virus sequences. On the left, a phylogenetic tree representing the large cluster is displayed. Collapsed branches consist of *Tombusviridae*, tombus-like viruses, and *Nodaviridae* capsid genes.

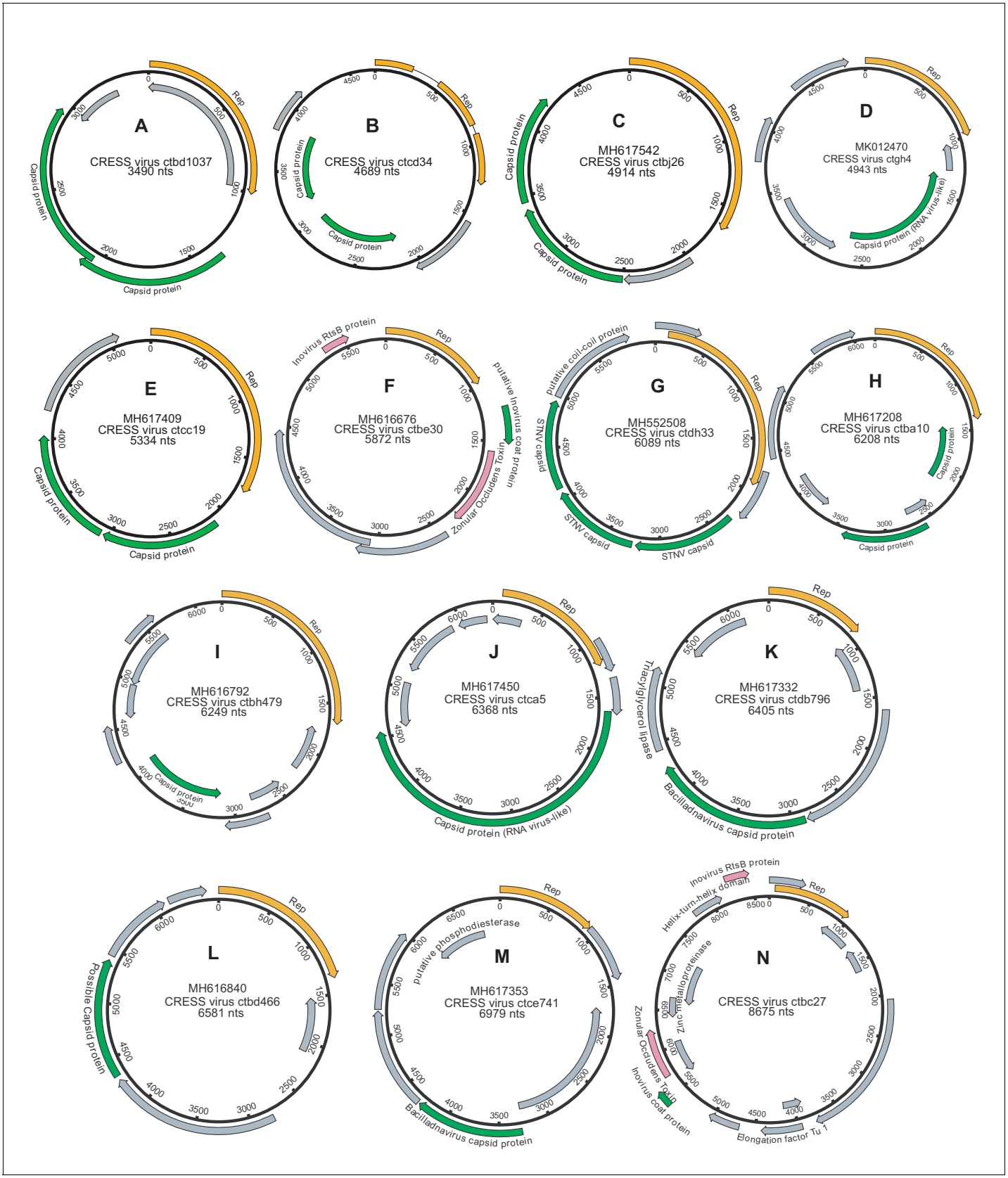
A
CRESS virus ctbd1037
3490 nts
Rep
Capsid protein
Capsid protein
B
CRESS virus ctcd34
4689 nts
Rep
Capsid protein
Capsid protein
C
MH617542
CRESS virus ctbj26
4914 nts
Rep
Capsid protein
Capsid protein
D
MK012470
CRESS virus ctgh4
4943 nts
Rep
Capsid protein (RNA virus-like)
E
MH617409
CRESS virus ctcc19
5334 nts
Rep
Capsid protein
Capsid protein
F
MH616676
CRESS virus ctbe30
5872 nts
Inovirus RtsB protein
Rep
putative Inovirus coat protein
Zonular Occludens Toxin
G
MH552508
CRESS virus ctdh33
6089 nts
putative coil-coil protein
Rep
STNV capsid
STNV capsid
STNV capsid
H
MH617208
CRESS virus ctba10
6208 nts
Rep
Capsid protein
Capsid protein
I
MH616792
CRESS virus ctbh479
6249 nts
Rep
Capsid protein
J
MH617450
CRESS virus ctca5
6368 nts
Rep
Capsid protein (RNA virus-like)
K
MH617332
CRESS virus ctdb796
6405 nts
Rep
Triacylglycerol lipase
Bacilladnavirus capsid protein
L
MH616840
CRESS virus ctbd466
6581 nts
Rep
Possible Capsid protein
M
MH617353
CRESS virus ctce741
6979 nts
Rep
putative phosphodiesterase
Bacilladnavirus capsid protein
N
CRESS virus ctbc27
8675 nts
Inovirus RtsB protein
Helix-turn-helix domain
Rep
Zn2+
metalloproteinase
Zonular Occludens Toxin
Inovirus coat protein
Elongation factor Tu 1

Figure 4—figure supplement 1. Genome maps of large CRESS virus genomes. Predicted CRESS Rep-like genes are displayed in orange, virion structural genes shown in green, other identifiable viral genes shown in pink, other genes in gray. GenBank accession numbers are displayed above the virus name.

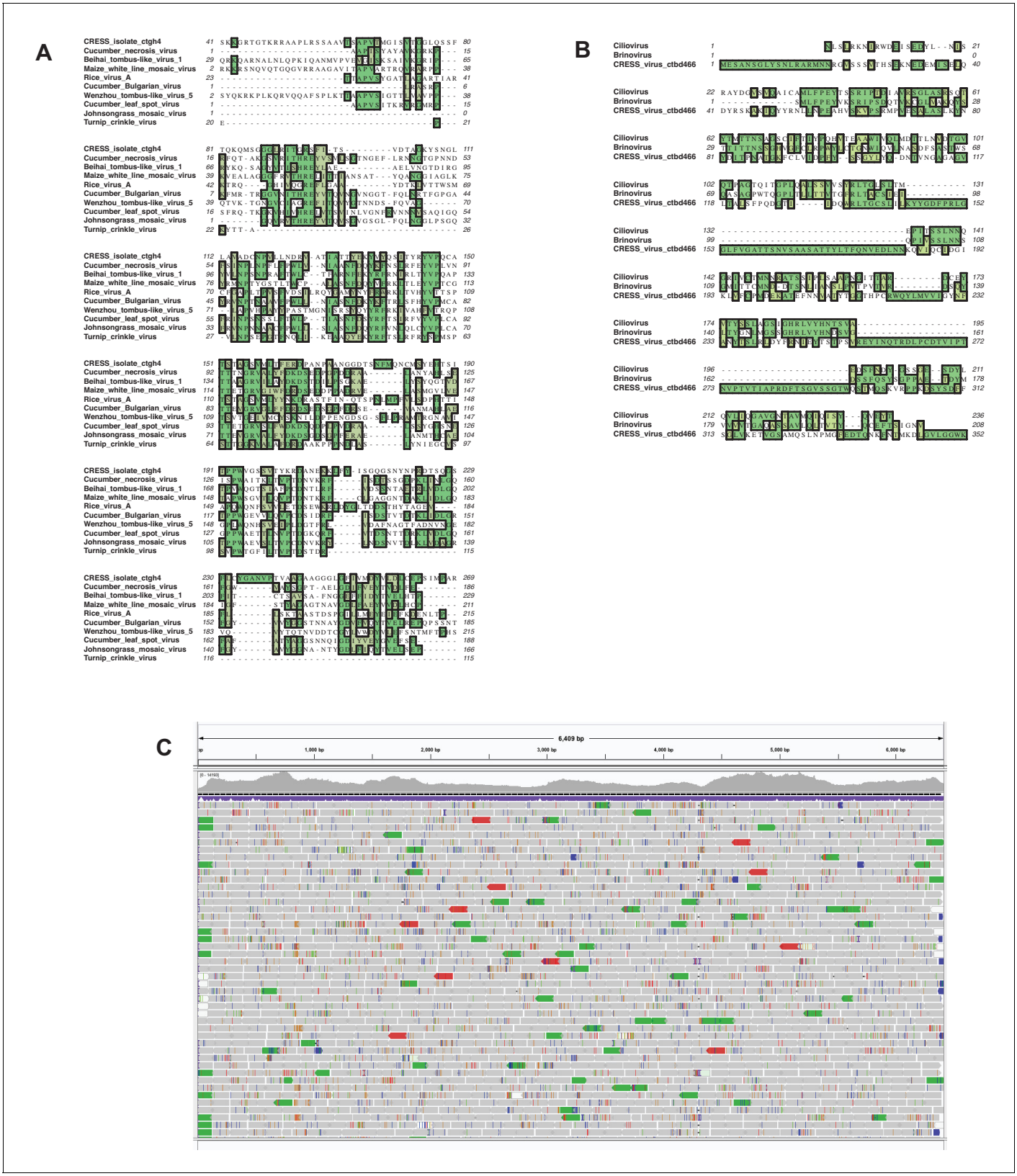

Figure 4—figure supplement 2. Validation of proteins with predicted similarity to RNA virus capsid proteins. (**A**) First order neighbors for Crucian-associated CRESS virus ctgh4 capsid protein were extracted from the network shown in ***Figure 5*** and aligned using Muscle. (**B**) The same approach was applied to CRESS virus ctbd466 capsid protein. (**C**) A visualization (Integrative Genomics Viewer) of a read alignment to CRESS virus isolate ctca5. The visualization shows no evidence of artifactual chimerization in the contig assembly process.

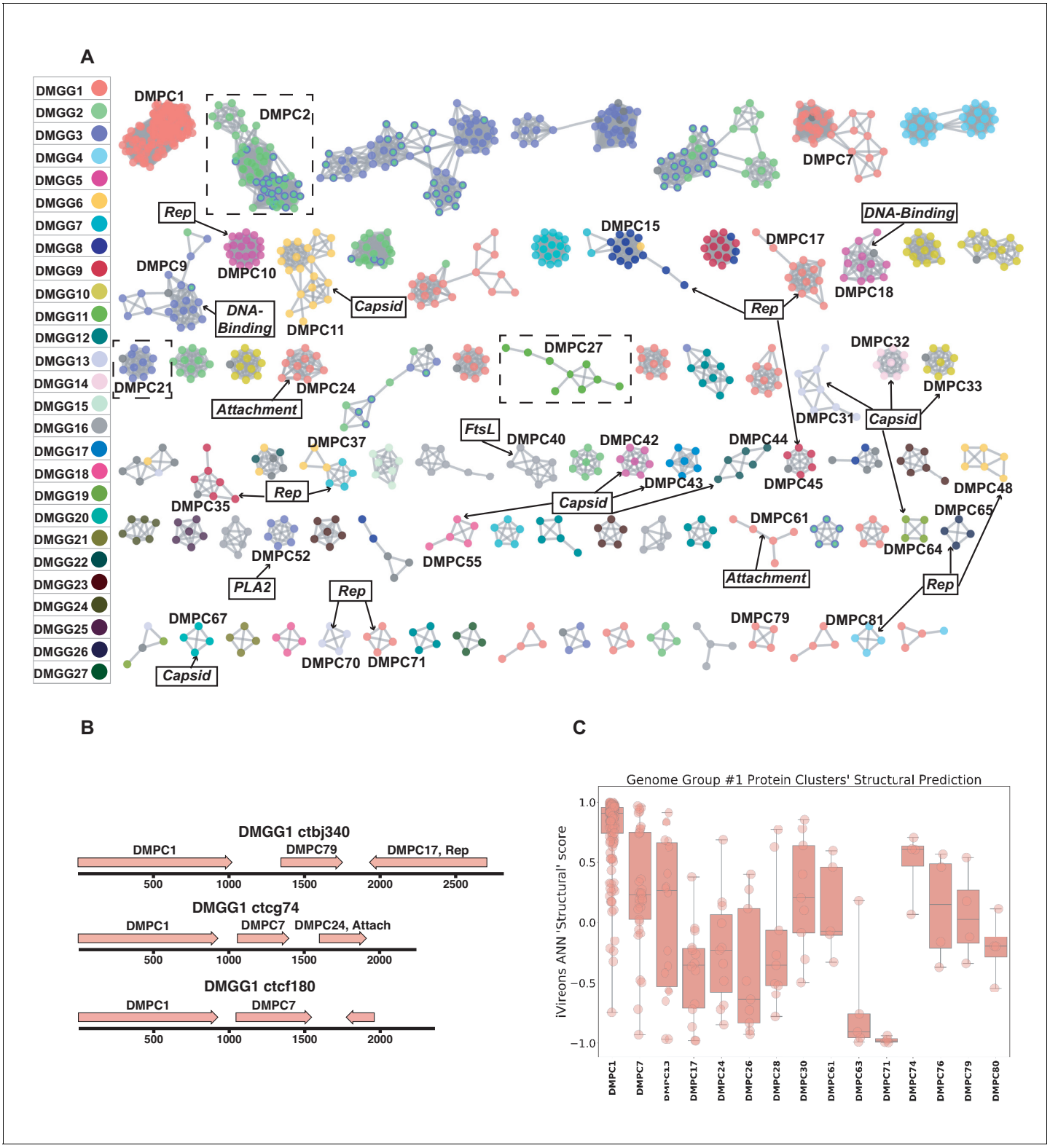

Figure 5. Dark matter analysis. (**A**) Sequence similarity network analysis for genes from dark matter circular sequences (minimum cluster size = 4). Clusters are colored based on assigned dark matter genome group (DMGG). Structural predictions from HHpred are indicated (>85% probability). *Rep* = rolling circle replicases typical of CRESS viruses or ssDNA plasmids. *Capsid* = single jellyroll capsid protein. *Attachment* = cell attachment proteins typical of inoviruses. *DNA-Binding* = DNA binding domain. *PLA2* = phospholipase A2. *FtsL* = FtsL like cell division protein. Clusters that contain a representative protein that was successfully expressed as a virus-like particle are outlined by a dashed rectangle (See ***Figure 6***). (**B**) Maps of

Figure 5 continued on next page

Figure 5 continued

three examples of DMGG1 with DMPCs labeled (linearized for display). (**C**) DMGG1 iVireons 'structure' score summary by protein cluster. Scores range from −1 (unlikely to be a virion structural protein) to 1 (likely to be a virion structural protein). Additional iVireons score summaries can be found in ***Figure 5—figure supplement 2***.

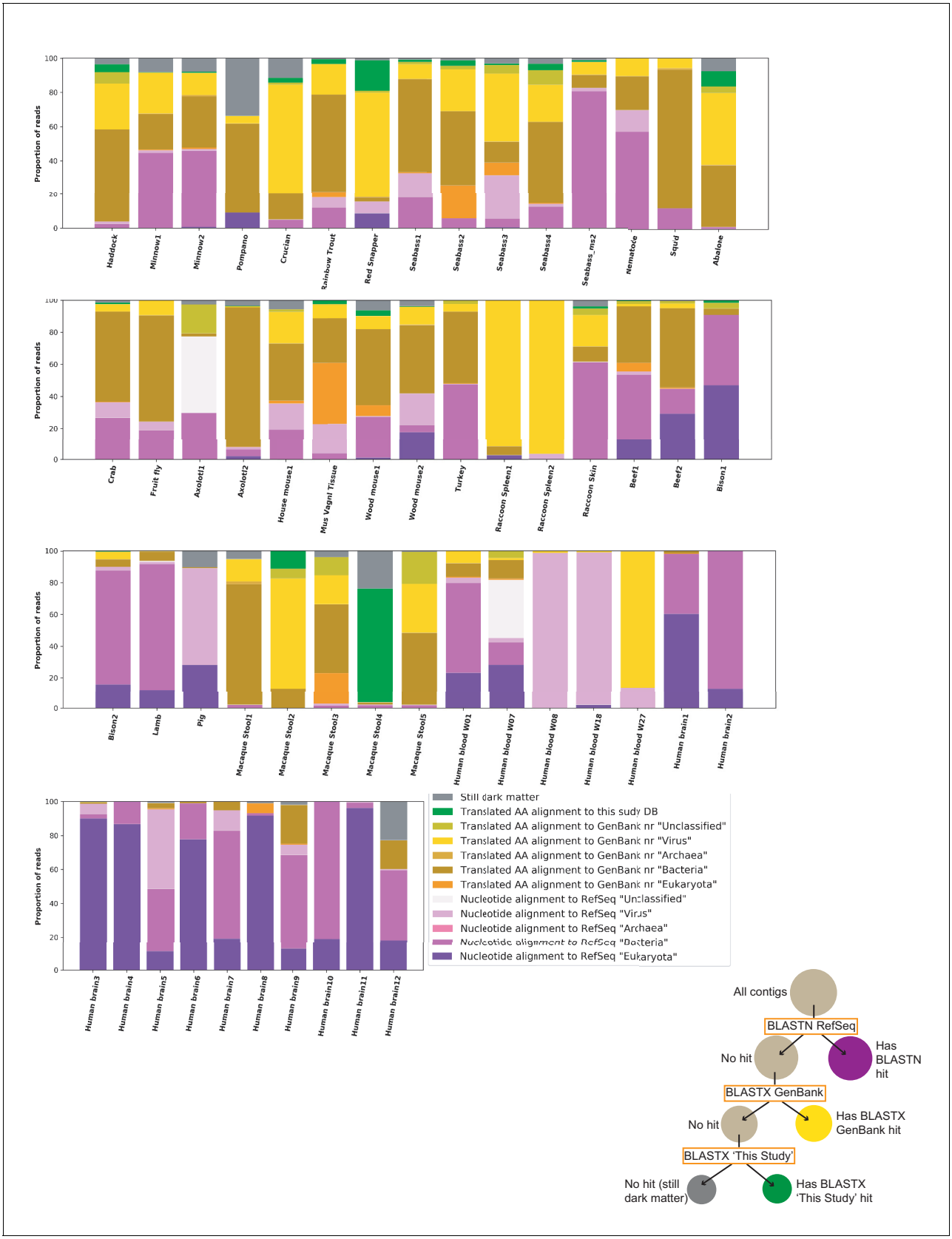

Figure 5—figure supplement 1. Sample characterization by iterative BLAST Searches. Contigs of over 1000 nts from each sample were subject to iterative BLAST searches. First, BLASTN was performed against the RefSeq database. Contigs without hits were then queried by BLASTX against all of

Figure 5—figure supplement 1 continued on next page

Figure 5—figure supplement 1 continued

GenBank 'nr' database. Contigs without hits were then queried by BLASTX against a database of proteins from genomes reported in this study. The proportion of total reads mapping to each contig was calculated and used for this plot. Individual inspection of contigs shows that most hits in the 'Translated AA alignment to GenBank' nr' 'Bacteria'' were likely plasmid or prophage proteins. The proportions of hits in each category are sensitive to stringency settings and to which databases are chosen for the analysis. The key aims of the figure are to display the proportion of reads the current survey rendered classifiable and the fraction of remaining dark matter reads in various samples.

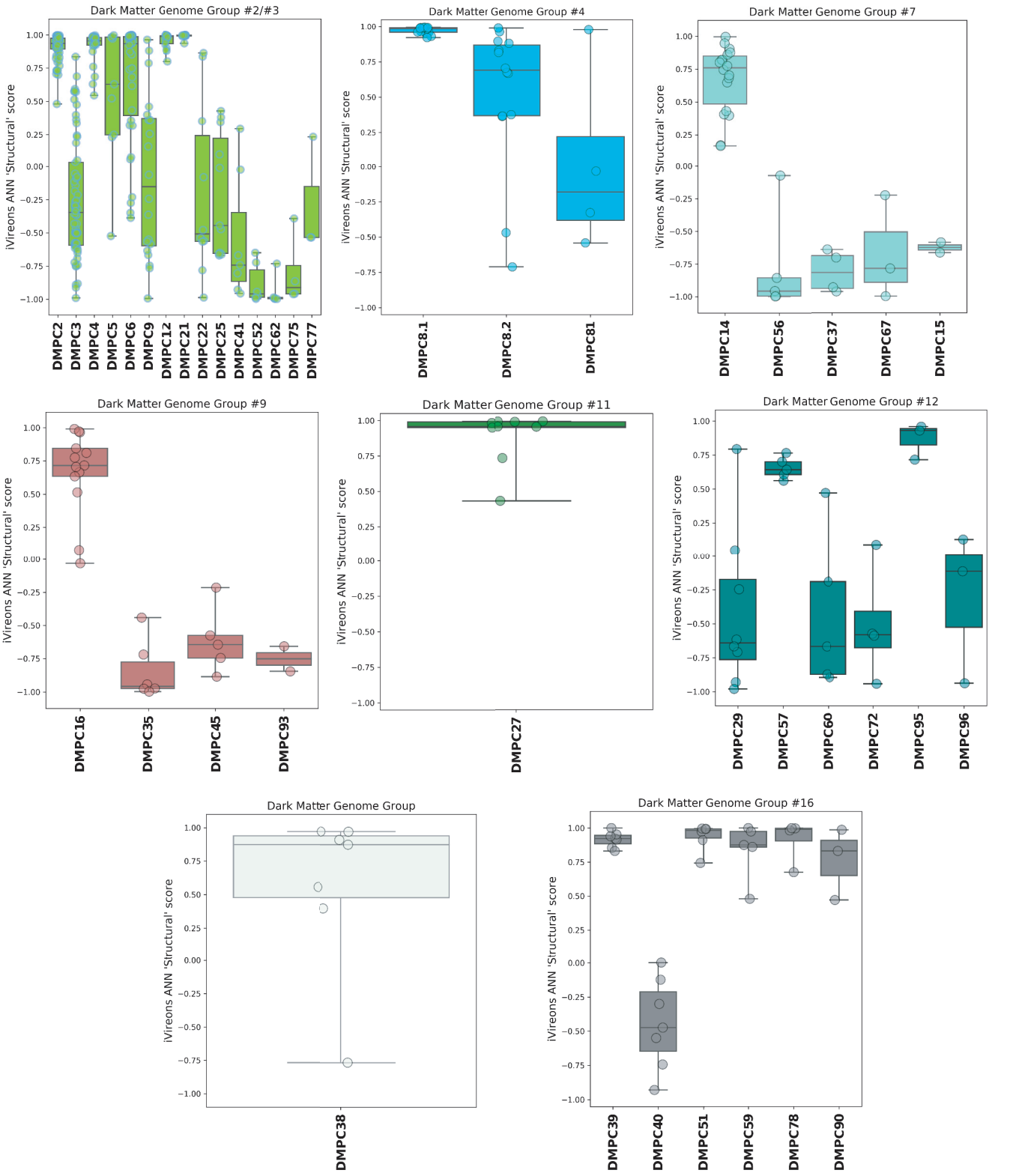
Dark Matter Genome Group #2/#3
iVireons ANN 'Structural' score
DMPC2
DMPC3
DMPC4
DMPC5
DMPC6
DMPC9
DMPC12
DMPC21
DMPC22
DMPC25
DMPC41
DMPC52
DMPC62
DMPC75
DMPC77
Dark Matter Genome Group #4
DMPC8.1
DMPC8.2
DMPC81
Dark Matter Genome Group #7
DMPC14
DMPC56
DMPC37
DMPC67
DMPC15
Dark Matter Genome Group #9
DMPC16
DMPC35
DMPC45
DMPC93
Dark Matter Genome Group #11
DMPC27
Dark Matter Genome Group #12
DMPC29
DMPC57
DMPC60
DMPC72
DMPC95
DMPC96
Dark Matter Genome Group
DMPC38
Dark Matter Genome Group #16
DMPC39
DMPC40
DMPC51
DMPC59
DMPC78
DMPC90

Figure 5—figure supplement 2. iVireons scores of DMGGs with candidate viral structural gene(s). Box-and-whisker plots of iVireons 'Structural' scores for individual DMPCs (numbers on x-axes) grouped by DMGG. Scores (y-axes) range from −1 (unlikely to be a virion structural protein) to 1 (likely to be a virion structural protein). DMGG2 and DMGG3 have been combined due to inferred chimerism.

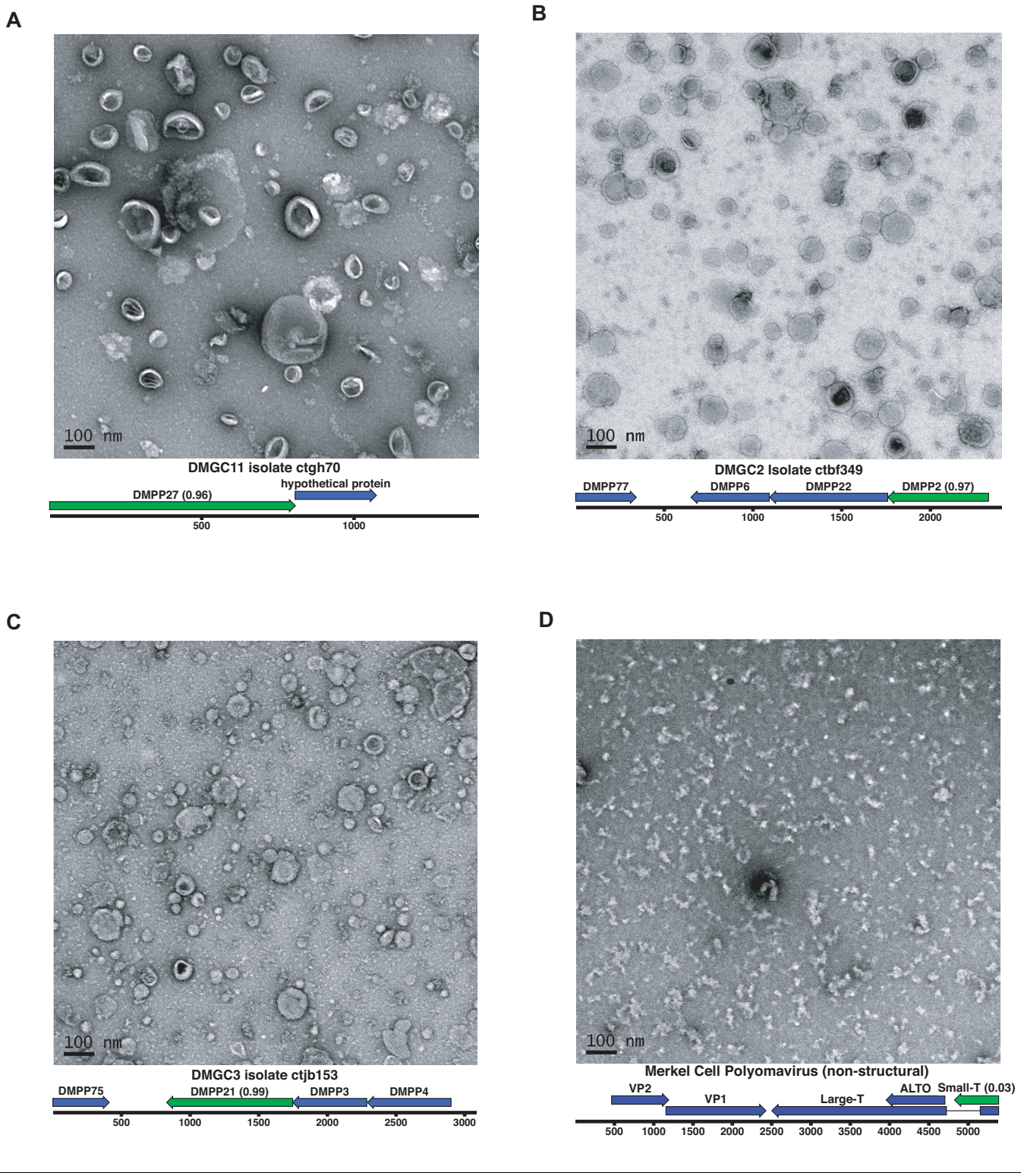

Figure 6. Expression of putative capsid proteins Images taken by negative stain electron microscopy. Genome maps are linearized for display purposes. Expressed genes are colored green. iVireons scores are listed in parentheses. (**A-C**) Images represent virus-like particles from iVireons-predicted viral structural genes. (**D**) Merkel cell polyomavirus small T antigen (a viral non-structural protein) is shown as a negative control.